UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

| | |
|---|---|
| RICHARD KADREY, et al.,<br>Plaintiffs,<br>v.<br>META PLATFORMS, INC.,<br>Defendant. | Case No. 23-cv-03417-VC (TSH)<br>**PUBLIC VERSION OF DISCOVERY ORDER AT ECF NO. 374**<br>Re: Dkt. Nos. 308, 361 |

Plaintiffs' RFP 118 requested "[a]ll Documents and Communications, including source code, relating to any efforts, attempts, or measures implemented by Meta to prevent Llama Models from emitting or outputting copyrighted material." In the joint discovery letter brief at ECF No. 308, Plaintiffs moved to compel the data mentioned in Sections 3 and 4.2 of Meta's Llama 2 paper and the data mentioned in Sections 4.2 and 5.4.3 of Meta's Llama 3 paper. That appeared to be a staggering amount of data, and at the hearing Plaintiffs made clear they were not seeking all of it. The Court ordered the parties to file a supplemental letter brief concerning this RFP, which the parties have now done. ECF No. 361. The Court held a hearing on January 8, 2025, and now issues the following order.

Plaintiffs now ask for four things. First, they seek the supervised fine-tuning data that Nikolay Bashlykov testified about at pages 144-46 of his deposition. It is apparently on a specific hard drive cluster referred to as EAG-WSF. Second, they request the post-training datasets used to train and fine-tune the Llama models specifically in reference to the "Intellectual Property" safety category. They say that this type of training is variously referred to as fine-tuning, safety fine-tuning, mitigation training, or similar terms. Third, Plaintiffs seek the post-training datasets comprising books sourced from the at-issue shadow datasets that are used for other safety categories. Plaintiffs clarify that they are not seeking all safety-related datasets, just the ones containing books from shadow libraries. Fourth, Plaintiffs request any additional post-training datasets sourced from shadow datasets and used by Meta to fine-tune its Llama models to minimize their ability to memorize or output training data verbatim. As to all four categories of data, Plaintiffs seek both the raw/original data from which these post-training datasets were created, as well as the data as specifically formed or constituted for use in the aforementioned post-training of the Llama models.

Let's start with the second category, the post-training datasets used to train and fine-tune the Llama models specifically in reference to the "Intellectual Property" safety category. The Court thinks these datasets are responsive to RFP 118. It's true that RFP 118 did not use the words "training data," but it did request "[a]ll Documents and Communications, including source code, relating to any efforts, attempts, or measures implemented by Meta to prevent Llama Models from emitting or outputting copyrighted material." The requested datasets are certainly "documents," and they seem to relate to efforts or attempts by Meta to prevent the Llama models from outputting copyrighted material. Section 5.4.7 of the Llama 3 paper makes clear that supervised fine-tuning was an effort to avoid intellectual property violations, and emitting or outputting copyrighted material seems like a type of intellectual property violation.

The Court also thinks Plaintiffs have made a sufficient showing of relevance. Plaintiffs' copyright claim is about Meta's use of their copyrighted materials to train the Llama models. Plaintiffs allege that "Meta made copies of the Infringed Works during the training process of the Llama 1 and Llama 2 language models without Plaintiffs' permission." ECF No. 133 ¶ 81. The Llama 2 and 3 papers make clear that fine-tuning is part of the training process for the Llama models. Here, a big factual dispute between the parties is whether the fine-tuning data consists of the copyrighted works themselves. Plaintiffs strenuously argue that it does, and Meta denies this. Both sides point to excerpts of Bashlykov's deposition to support their arguments.

The Court has reviewed the cited portions of this deposition (ECF Nos. 362-5 and 358-17) and finds the testimony ambiguous. Bashlykov was clear that parts of datasets that presumably contain copyrighted works were *used* in post-training and fine-tuning. He said the fine-tuning data was "based on" the dataset (page 146) and that parts of the dataset "were used as a kind of source file[]" (pages 71-72) to prepare the fine-tuning datasets. However, the cited testimony is not clear whether parts of datasets that contain copyrighted works are *in* the datasets used for fine-tuning. If the answer is yes, then the fine-tuning datasets are additional copies of the allegedly infringed works that were made during the Llama training process (granted, in post-training, not in pre-training, but still part of the overall effort to train), so come within the scope of Plaintiffs' copyright claim. If the answer is no, then it seems that the fine-tuning datasets are not relevant because they are neither infringing copies nor derivative works. Of course, Plaintiffs cannot definitively prove that the fine-tuning datasets contain infringing works because they don't have them. The Court is mindful that Plaintiffs are not required to prove their case on the merits in order to obtain discovery. Rather, Plaintiffs' burden on a motion to compel is to show that the requested discovery is a worthwhile endeavor in view all of the factors in Rule 26(b)(1). Here, Plaintiffs have made a sufficient factual showing that the use of datasets that contain copyrighted works to create datasets that were used in fine-tuning the Llama models concerning intellectual property violations may have or could have resulted in portions of the copyrighted works ending up in the fine-tuning datasets, such that Plaintiffs are entitled to learn if that in fact happened.

As a back up argument, Meta argues that even if the fine-tuning datasets contain additional copies of copyrighted works, it is unnecessary to produce those datasets because the additional copies would just be a subset of the pre-training datasets that Meta has already produced. Based on the cited portions of Bashlykov's deposition, Meta's subset argument appears to be factually correct. Nonetheless, if the fine-tuning datasets contain additional copies of copyrighted works, those are additional allegedly infringing copies that were used to perform a somewhat different training task. Plaintiffs are allowed to take discovery into the full scope of their copyright claim, not just part or most of their claim. The Court previously denied Plaintiffs discovery into all copies of copyrighted works both because the RFP at issue did not request that and because "this case . . . is about the use of copyrighted materials to train the Llama models, not all copyright infringement committed by Meta." ECF No. 351 at 2. Plaintiffs' current request for fine-tuning datasets related to intellectual property is within the scope of their claim. Accordingly, the Court will grant that portion of Plaintiffs' motion to compel.

Plaintiffs' third category is the post-training datasets comprising books sourced from the at-issue shadow datasets that are used for other safety categories. This category suffers from two problems. First, the Court has already explained that "shadow" datasets is a pejorative description of a dataset, and Meta can't be expected to guess what Plaintiffs think falls within that term. ECF No. 315 at 7-8. Second, according to section 5.4.7 of the Llama 3 paper, the "other" safety categories are child sexual exploitation, defamation, elections, hate, indiscriminate weapons, non-violent crimes, privacy, sex-related crimes, sexual content, specialized advice, suicide and self-harm, and violent crimes. None of those topics have anything to do with RFP 118.

Plaintiffs' fourth category is any additional post-training datasets sourced from shadow datasets and used by Meta to fine-tune its Llama models to minimize their ability to memorize or output training data verbatim. This category is vague. We again have the "shadow" datasets problem. The Court discussed this fourth category with the parties at the January 8 hearing, and the Court is not satisfied that this category refers to anything in particular. Plaintiffs made clear during the hearing that this category as drafted is not limited to things that are relevant to this lawsuit because the "training data" Llama models are trained not to output includes personally identifying information – an entirely separate concern from copyright infringement. The Court agrees with Meta that the purpose of having the parties file a supplemental letter brief on RFP 118 was for Plaintiffs to refine their previous overbroad request and for Meta to be able to respond to that refinement. With respect to category four, Plaintiffs have not done that. The Court will therefore not compel category four.

Now let's go back to the first category, the SFT data identified by Bashlykov. There is no evidence before the Court concerning whether this data was used to fine-tune for the intellectual property safety classification. If it was, it's relevant and responsive. But if it was used only to fine-tune for other safety categories, then it's not responsive to RFP 118, which is about preventing the Llama models from emitting or outputting copyrighted material.

Finally, we have Plaintiffs' request for both the raw/original data from which the post-training datasets were created, as well as the data as specifically formed or constituted for use in the post-training of the Llama models. However, Section 3.1 of the Llama 2 paper suggests that the raw or original data is not only massive compared to the datasets actually used, but also of

United States District Court
Northern District of California

dubious value:

> **Quality Is All You Need.** Third-party SFT data is available from many different sources, but we found that many of these have insufficient diversity and quality—in particular for aligning LLMs towards dialogue-style instructions. As a result, we focused first on collecting several thousand examples of high-quality SFT data, as illustrated in Table 5. By setting aside millions of examples from third-party datasets and using fewer but higher-quality examples from our own vendor-based annotation efforts, our results notably improved. These findings are similar in spirit to Zhou et al. (2023), which also finds that a limited set of clean instruction-tuning data can be sufficient to reach a high level of quality. We found that SFT annotations in the order of tens of thousands was enough to achieve a high-quality result. We stopped annotating SFT after collecting a total of 27,540 annotations. Note that we do not include any Meta user data.

Section 5.4.7 of the Llama 3 paper also states that Meta did "extensive cleaning" of collected samples to improve the performance of Llama Guard 3. The Court therefore concludes that the raw or original data from which the post-training datasets were created is not proportional to the needs of the case.

Accordingly, the Court **GRANTS IN PART** and **DENIES IN PART** Plaintiffs' motion to compel. The Court **ORDERS** Meta to produce the post-training datasets used to train and fine-tune the Llama models specifically in reference to the "Intellectual Property" safety category. The Court also **ORDERS** Meta to produce the SFT data identified by Bashlykov if it was used to fine-tune one or more Llama models for the intellectual property safety category. The Court **ORDERS** Meta to serve a declaration on Plaintiffs within seven days stating whether that SFT data was used for that purpose or not. The Court otherwise **DENIES** Plaintiffs' motion to compel.

**IT IS SO ORDERED.**

Dated: January 8, 2025

THOMAS S. HIXSON
United States Magistrate Judge